Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Knowing when to fold 'em: Problem attributes and strategy differences in the Paper Folding test

Heather Burte[a,*], Aaron L. Gardony[a,b,c], Allyson Hutton[d], Holly A. Taylor[a]

[a] *Tufts University, Department of Psychology, 490 Boston Ave, Medford, MA 02155, USA*
[b] *Center for Applied Brain & Cognitive Sciences, 200 Boston Ave, Medford, MA 02155, USA*
[c] *Cognitive Science Team, U.S. Army Natick Soldier Research, Development, and Engineering Center, Natick, MA 01760, USA*
[d] *Think3d!, 1018 Maple Lane, Davis, CA 95616, USA*

ARTICLE INFO

Keywords:
Spatial visualization
Spatial skills
Psychometric measures
Strategies

ABSTRACT

Spatial visualization—the ability to mentally imagine and manipulate objects—has frequently been measured using the Paper Folding Test (PFT). In this task, participants view diagrams of a paper being folded and a hole being punched. They then identify the resulting punch pattern. Although task instructions promote mentally unfolding the paper, the extent to which people follow this spatial visualization strategy is unknown. The present work assesses hypothesized PFT solution strategies and their relation to problem attributes. Accuracy was impacted by the interaction between fold types, linear mixed models revealed greater use of simple heuristics compared to the suggested unfolding. Furthermore, most participants used a single strategy but simple heuristics were more often used than unfolding. Given this, we recommend redesigning the PFT to utilize the prevalence of strategy use to assess individual differences.

1. Introduction

People vary in spatial visualization skills, frequently measured using the Paper Folding Test (PFT; Ekstrom, French, Harman, & Dermen, 1976). In the PFT, participants view diagrams of a paper being folded and then a hole punched through it. They then identify the spatial arrangement of punches after unfolding. This test purportedly evaluates mental manipulation or "spatial visualization" skills involved in mentally unfolding the paper. Spatial visualization is separable from other spatial skills (Hegarty & Waller, 2004; Lohman, 1979).

Despite the assumption that the PFT measures spatial visualization, few studies have explored potential strategies and/or the role problem attributes may play. Cognitively, spatial visualization, memory span, and visual memory are related to PFT strategy use and performance (Kyllonen, Lohman, & Snow, 1984). Although most participants reported mentally unfolding, performance reflects multiple strategies (Jaeger, 2015). Further, the few studies that have explored problem attributes found that difficulty increased with the number, types, and combination of folds (Jaeger, 2015; Wothke & Zimowski, 1988). It remains unclear whether all participants use spatial visualization and/or whether they use other strategies. Answers to these points would indicate the extent to which PFT assesses spatial visualization. Here, we investigated how problem attributes impacted PFT accuracy and signaled the use of four potential strategies.

1.1. Connecting problem attributes and strategy differences

Commonly used spatial tasks (e.g., mental rotation) purportedly assess a single cognitive approach, yet multiple approaches seem likely. Using a task similar to the Paper Folding test, Shepard and Feng (1972) found that task and stimulus factors influenced performance. Just and Carpenter (1985) found that the assumption that people mentally rotate around a common axis did not sufficiently describe strategies evident in cube comparisons (Ekstrom et al., 1976) or mental rotation (Shepard & Metzler, 1971). Just and Carpenter then collected self-reported strategies and developed problem-level predictions based on them. Problem attributes, such as trajectory complexity or rotational difference between figures, invoked different strategies on different problems. This finding echoes Lohman's (1979) conclusion that difficult problems (defined by problem characteristics and required mental operations) elicit more strategy types than easy problems. Two strategies typical in visual comparison tasks involve holistic (comparing the entire figure at once) and analytic (comparing based on figure components; Cooper, 1976) processing. This distinction has been applied when

[*] Corresponding author.
E-mail addresses: heather.burte@tufts.edu (H. Burte), aaron.gardony.civ@mail.mil (A.L. Gardony), ahutton.think3d@gmail.com (A. Hutton), holly.taylor@tufts.edu (H.A. Taylor).

https://doi.org/10.1016/j.paid.2018.08.009
Received 31 October 2017; Received in revised form 14 August 2018; Accepted 16 August 2018
0191-8869/ © 2018 Elsevier Ltd. All rights reserved.

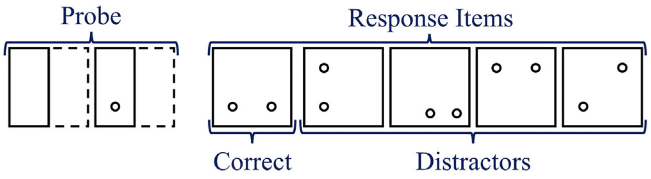

Fig. 1. An example PFT problem with its probe and response items.

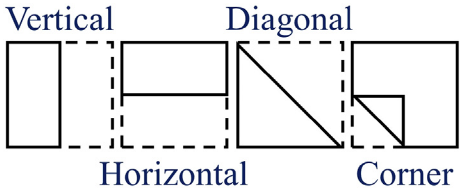

Fig. 2. Four PFT fold types.

understanding individual differences in mental rotation performance (Khooshabeh, Hegarty, & Shipley, 2013; Li & O'Boyle, 2013).

Overall success on spatial tasks may increase when flexibly employing different strategies, depending on individual problems. In fact, mental rotation strategies change for same versus different trials (Gardony, Taylor, & Brunyé, 2014), and shift as problem difficulty increases (Gardony, Eddy, Brunyé, & Taylor, 2017). These studies illustrate why investigating strategies is essential, and that problem attributes and strategies often interact.

In this paper, we investigated how problem attributes both impacted PFT accuracy and signaled use of four potential strategies. Previous work suggested that difficulty increases with number of folds, occlusions, asymmetrical folds, and fold combinations (Jaeger, 2015; Kyllonen et al., 1984). Beyond these studies, research has not systematically analyzed how PFT problem attributes impact strategy use. Knowing the answer to this would help interpret individual differences in PFT accuracy and its relationship to other spatial measures. The present work used linear mixed models (LMM; Bates, Mächler, Bolker, & Walker, 2015) to predict accuracy and identify strategies based on problem attributes individually and nested together under each participant. Furthermore, a strategy categorization scheme was developed to provide insights into each participants' strategy use.

1.2. Problem attributes and predictions

Each PFT problem includes the *probe* and the *response items* (Fig. 1). The probe contains images depicting a paper being folded and a hole punched through it. Probes include four *fold types*: *horizontal*, *vertical*, *corner*, and *diagonal* (Fig. 2). The *punch locations* play critical roles in the response items. The response items depict possible punch configurations after unfolding the paper. The *correct item* accurately depicts punch locations, whereas the *distractor items* depict incorrect locations. Based on the probes, we identified problem attributes that might impact accuracy. With these in mind, strategies can then be identified based on the response item selected relative to the probe attributes.

We investigated three problem attributes: number of folds, type of fold, and occlusion (Table A.1). Although the PFT instructions ask participants to imagine unfolding the paper, some problems can be answered without unfolding all folds (Fig. 3). To distinguish between required and unrequired folds, the term *number of folds* indicates all folds and *number of relevant folds* indicates only folds necessary to identify the correct answer.

Although every fold occludes a part of the paper, most folds result in the edges of the occluded part being still visible, or the edges of the folds lie directly on top of one another (Fig. 4). In other instances, an “occlusion” occurs. Occlusions happen when a fold “hides” another fold, in a way that some of the occluded fold's edge is no longer visible. *Fold occlusions* occur when the location of the punch falls outside the occluded area whereas *punch occlusions* occur when the punch location falls within the occluded area.

We made a priori predictions about how each problem attribute would impact PFT accuracy based on the demands the attributes made on cognitive processing. Given that each additional fold likely necessitates additional cognitive effort, we hypothesized that accuracy should decrease with more folds. If while viewing the probe, people identify which folds contribute to solving the problem, they may focus only on relevant folds. If so, PFT accuracy should decrease with more relevant folds as long as assessing a fold's relevance does not carry additional cognitive effort.

Whereas single folds may be easy to visualize, fold combinations may be more challenging. We hypothesized that problems involving occlusion (particularly punch occlusion) require more cognitive effort due to occluded details that must be held in memory. Given this, we predicted that horizontal and vertical folds will be easiest because occlusions are rare and there are limited possible fold combinations. Corner folds (in combination with horizontal and vertical folds) are moderately difficult by introducing more complex fold combinations. Finally, diagonal folds (in combination with other folds) present the biggest challenge by introducing complex combinations that routinely include occlusions. The interaction of diagonal folds with other fold types also produces “non-perceptual matches”. These are problems where the punch location in the probe (e.g., in the top-right corner) does not match the punch location in the correct answer (e.g., in the bottom-right corner). We hypothesized that not having a match between probe and response punch locations also requires more effort due to memory demands.

1.3. Strategies and predictions

We identified four potential PFT strategies based the PFT instructions and insights from participants during debriefing sessions in previous studies. Using the relationship between probe items and the likelihood of incorrectly selecting distractors items (Appendix A), we sought to determine whether PFT accuracy could reveal strategy use.

The PFT instructs participants to use a **folding-unfolding strategy**, namely to imagine unfolding the paper to reveal the punch

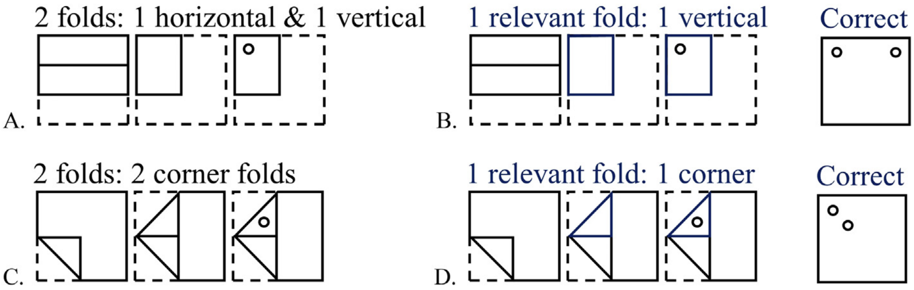
2 folds: 1 horizontal & 1 vertical
A.
1 relevant fold: 1 vertical
B.
Correct
2 folds: 2 corner folds
C.
1 relevant fold: 1 corner
D.
Correct

Fig. 3. Two methods of counting folds with correct answers on the right. A and C shows all of the folds, then using the same probes, B and D shows relevant folds.

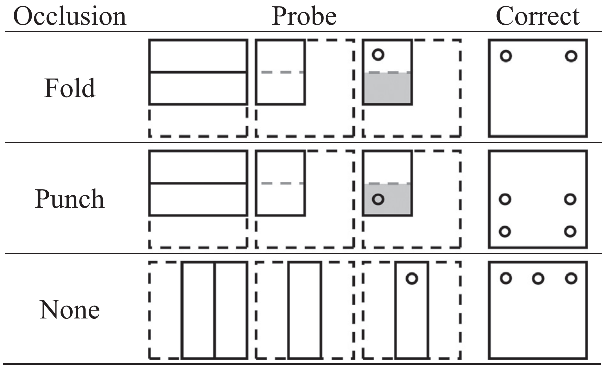

Fig. 4. Probes depicting fold occlusion, punch occlusion, and no occlusion. The grey dashed line (not in the PFT) represents the occluded edge. The light grey fill (also not in the PFT) represents the occluded area.

configuration (Fig. 5). By explicitly instructing this strategy, it makes sense that the PFT is thought to assess spatial visualization. This strategy's high cognitive load might make it unattractive for complex problems or for individuals with poor spatial skills. When using this strategy, more folds should negatively impact accuracy, but it might depend on whether participants consider all or only relevant folds. Accuracy should also reflect the number of folds implicitly depicted in distractor items (Table A.2). For instance, a common error when using this strategy might be to skip mentally unfolding a fold. A successful distractor might depict the punch configuration resulting from this error. For example, in a three-fold problem, distractors that implicitly depict two folds might better "distract" from the correct item. In other words, when distractor items successfully "distract" then accuracy should drop.

The **perceptual match strategy** involves matching the punch location in the probe to the punch locations in the response items. If the probe shows a center-top punch (Fig. 6), a participant might select a response item with a center-top punch (i.e., perceptual match) and/or rule out items without a center-top punch (i.e., non-perceptual match). Although more efficient than the folding-unfolding strategy, the perceptual match strategy is more error-prone. A limitation of this strategy is that not all correct answers have a perceptual match. For 16 of 20 problems, the correct item has a perceptual match (Table A.3), so accuracy should be better on these problems. Accuracy will also depend how many distractors have a perceptual match. When more distractors have perceptual matches, accuracy should decrease.

Fold-to-punch algorithms rely on a simple rule: each fold yields two holes after a punch. Given this, a participant might simply count the number of folds, multiply by two, and select a response with that number of punches. There are two fold-to-punch algorithms: *simple* and *relevant*. The simple fold-to-punch algorithm counts all folds (Fig. 7) whereas the relevant fold-to-punch algorithm incorporates only relevant folds (Fig. 8). For nine of 20 problems, the correct item follows the simple algorithm (Table A.3) and for 14 out of 20 problems, the correct answer follows the relevant algorithm. Accuracy should be better when correct answers conform to the algorithm (simple or relevant). Accuracy should also relate to the number of distractors that conform to these algorithms. On three of the nine problems where the correct item follows the simple algorithm and on three of the 14 problems where the correct item follows the relevant algorithm, no distractors follow the algorithms. In these cases, following a fold-to-punch algorithm guarantees a correct answer. Although these strategies do not always lead to finding the correct answer, they can sometimes lead to the correct answer and/or they can reduce the set of response items.

To summarize, we investigated how PFT problem attributes (number of folds, type of folds, and occlusion) impact accuracy and how these attributes reveal evidence for strategy use (folding/unfolding, perceptual match, and simple or relevant fold-to-punch).

2. Methods

2.1. Participants

Participants were recruited from Amazon's Mechanical Turk (MTurk), an online platform that crowd sources "workers" to complete online tasks. Participants gave informed consent online before and were debriefed online after the experiment in accordance with the Declaration of Helsinki. MTurk workers with > 95% completion rates, fluent in English, and in the US completed the experiment. In total, 220 completed the experiment, but 17 were dropped due inability to pair data, 18 were dropped for PFT performing below chance (< 20%), seven were dropped for completing the PFT in under 2 min, and 29 were dropped for failing to properly answer a catch question. The remaining 149 (68%) participants (average age: 32 years; range: 19–64; *SD*: 8 years; 39% female) were included in the analysis.

2.2. Design

This exploratory work tests within-subject effects of problem attributes and hypothesized strategies on accuracy rates. To do this, we used a series LMMs that allow for an analysis of problem attributes and strategies, nested together under each participant.

2.3. Materials and procedure

Using PsiTurk (Gureckis et al., 2016), participants started by completing the PFT (Ekstrom et al., 1976), one problem at a time at their own pace. A 1 s fixation appeared between problems. Although participants completed other spatial measures and demographics questions, this manuscript focuses on the PFT. We redrew high resolution PFT problems, maintaining the features and look of the original PFT. Unlike the original PFT, problems appeared in random order within sets of 10 (problems 1–10; problems 11–20).

3. Results

3.1. Performance

The PFT is generally administered as a paper-and-pencil task, so performance generally equates to accuracy. Participants did well (mean = 12.7/20 or 64%; Fig. 9), but there were large individual differences (range = 5–20 correct; *SD* = 3.5). It should be noted that we also ran all of the reported models using reaction time data, and the reaction time models generally mimicked the accuracy results.

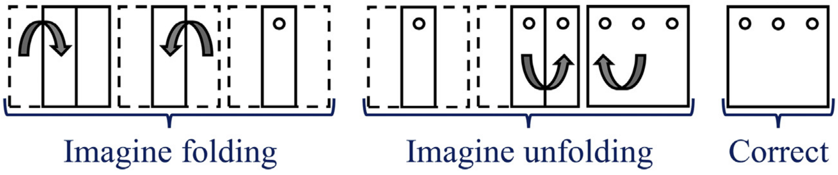

Fig. 5. Folding-unfolding: Participant imagines the probe folding (left, arrows added for emphasis). Participant imagines unfolding (center; not in the PFT). The correct answer emerges from correctly applying the folding-unfolding strategy (right).

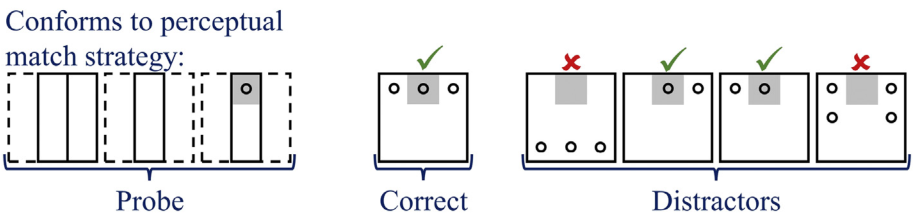

Fig. 6. Perceptual match: In the probe, the participant identifies the punch location (light grey for emphasis). The correct item has a perceptual match as do some distractors (indicated by checkmark). Other distractors do not (indicated by x).

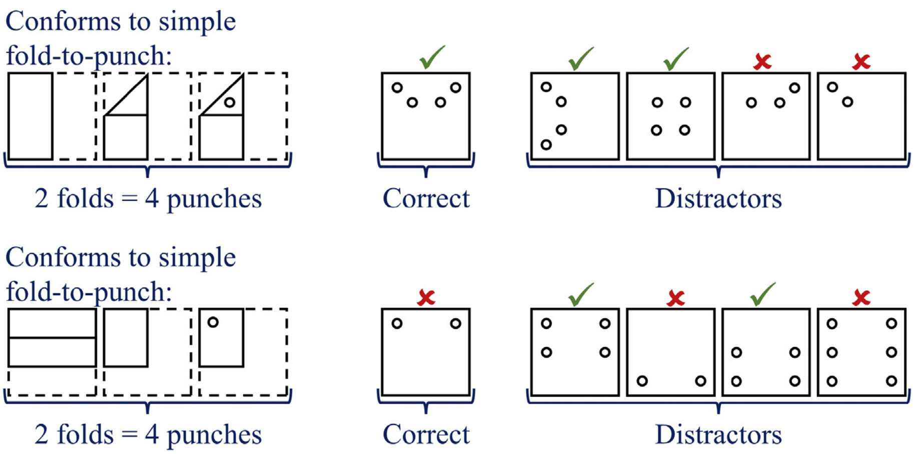

Fig. 7. Simple fold-to-punch: Top row, the probe has two folds, translating into four punches. The correct item and two distractor items have four punches (checkmarks). Two distractors do not (x). Bottom row, two probe folds translate into four punches. The correct answer does not have four punches.

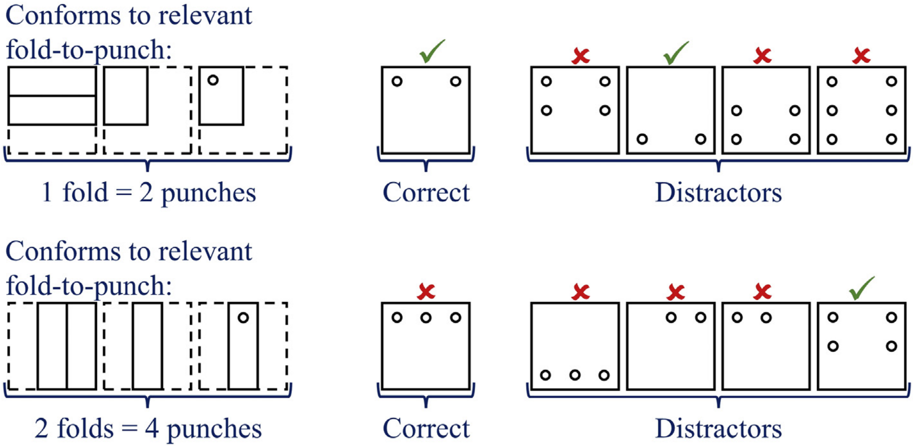

Fig. 8. Relevant fold-to-punch: Top row, of the two probe folds only the vertical fold is relevant. One relevant fold translates into two punches. The correct item and one distractor have two punches. Three distractors do not. Bottom row, two relevant folds translates into four punches. The correct item and three distractors do not have four punches.

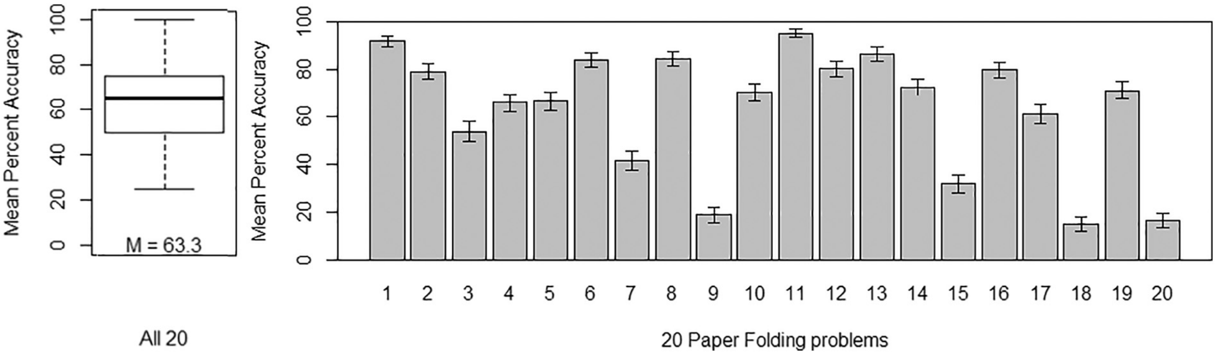

Fig. 9. Accuracy for all problems (left) and for each problem (right). For box plots, the box's center represents the median, the top and bottom indicate the first and third quartile, the whiskers indicate a 95% confidence interval, and mean values are labeled.

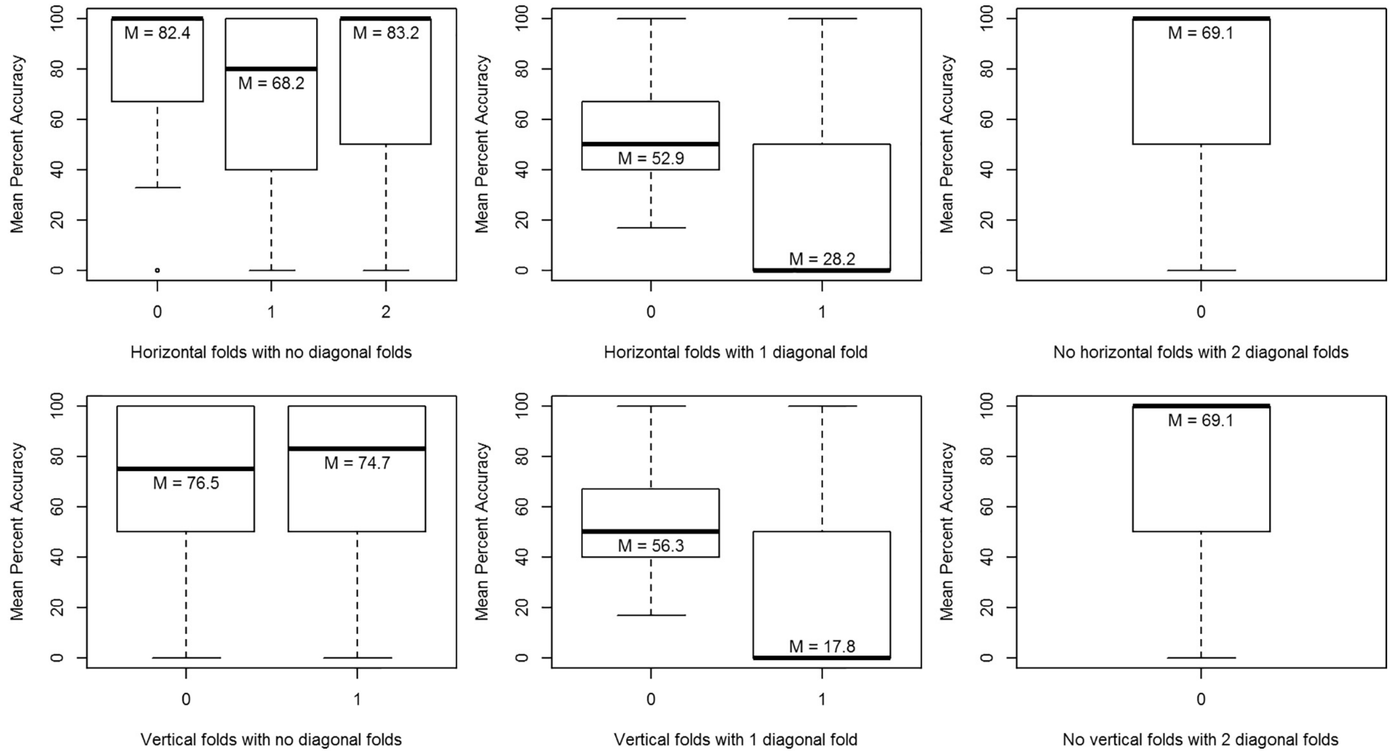

Fig. 10. Accuracy predicted by the interaction between horizontal and diagonal folds (top row) along with vertical and diagonal folds (bottom row).

3.2. Problem attributes

LMMs assessed how problem attributes predicted accuracy (see Appendix B). A model that included fold type interactions found that diagonal folds interacted with both horizontal and vertical folds, and this model outperformed the null model, $\chi^2(2) = 19.19, p < .001$ (Table B.1). Diagonal folds interacted with horizontal and vertical folds similarly (Fig. 10). The pattern suggests that mixed fold types yielded the lowest accuracy and that diagonal folds were more difficult than horizontal/vertical folds. The highest accuracy occurred when items had no diagonal folds with zero to two horizontal or vertical folds. Zero or two diagonal folds with no horizontal/vertical folds led to intermediate accuracy.

3.3. Strategy use

We developed five LMMs to investigate the four hypothesized strategies (Appendix B).

3.3.1. Folding-unfolding

For all folds, the number of folds (probe effect), number of distractors implicitly depicting fewer than all folds (distractor effect), and interaction were all significant (Figs. 11 and 12). This model outperformed the null model, $\chi^2(3) = 8.02, p < .05$ (Table B.2). Accuracy decreased with increasing folds, suggesting that participants may evaluate each fold's implication for the correct answer. Accuracy decreased with more distractors implicitly showing fewer than the total number of probe folds (i.e., folding-unfolding distractors). This finding provides evidence that some participants attempted to use the folding-unfolding strategy, but ineffectively applied it. They were drawn away from the correct answer by folding-unfolding distractor items.

The interaction between number of folds and folding-unfolding distractors revealed that, for one-fold problems, accuracy was high and unaffected by distractors. This is likely because it is impossible to have distractors that imply the unfolding one less fold in a one-fold problem. For two-fold problems, accuracy was slightly better when no folding-

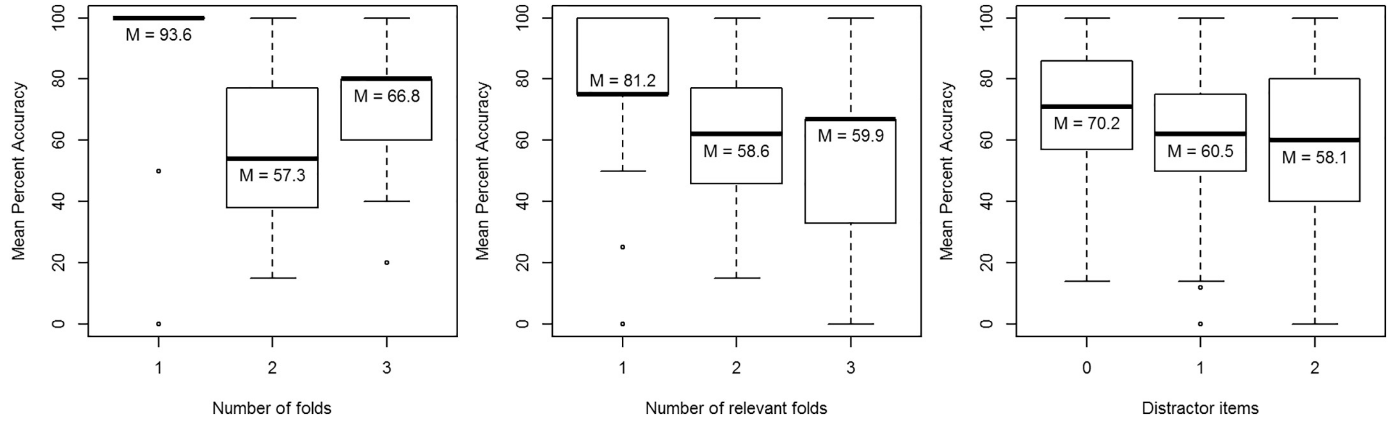

Fig. 11. Folding-unfolding: Accuracy predicted by number of folds (left), number of relevant folds (center), and number of distractor items depicting fewer than all folds (right).

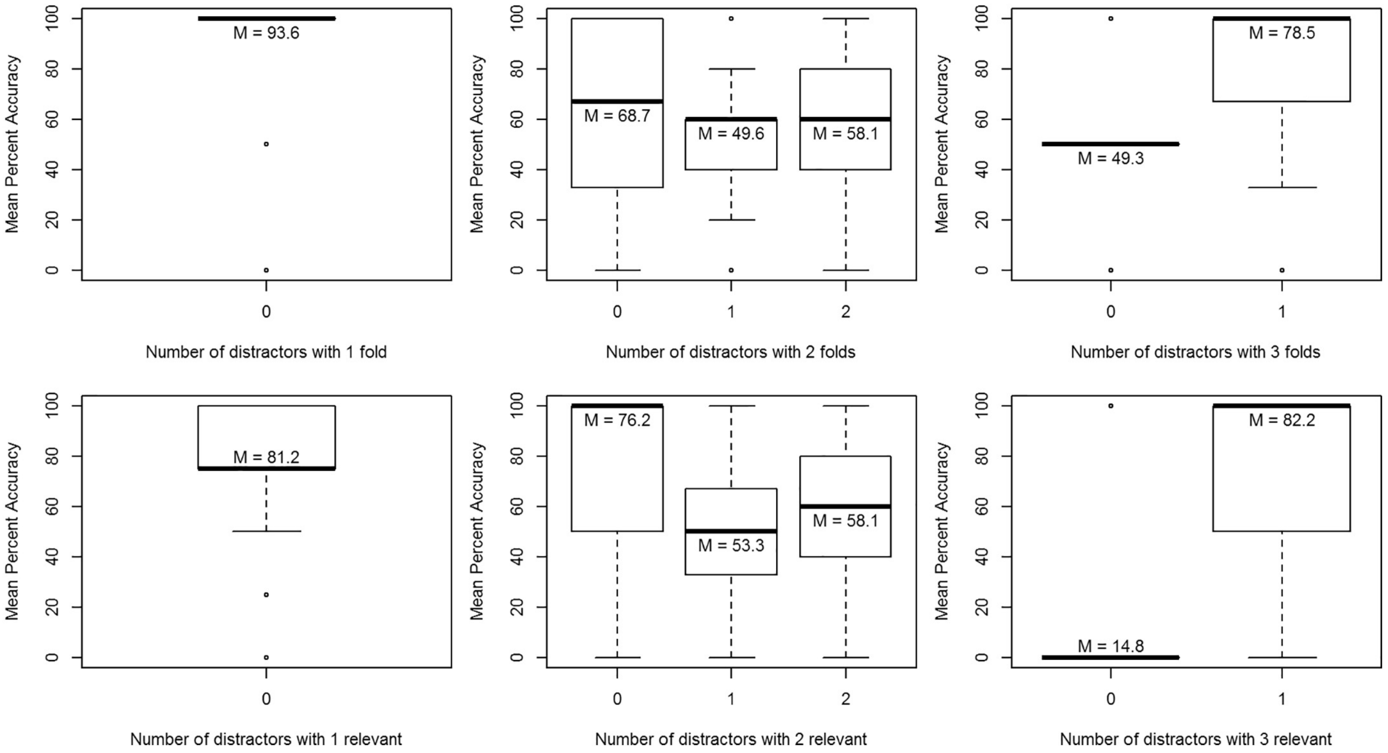

Fig. 12. Folding-unfolding: Accuracy predicted by the interaction of number of folds with number of distractor items depicting fewer than all folds (top row), and by the interaction of number of relevant folds with number of distractor items depicting fewer than all relevant folds (bottom row).

unfolding distractors existed. This indicates the strong draw of folding-unfolding distractors. For three-fold problems, something else must be driving accuracy rates as participants do better with one folding-unfolding distractor than with none. Using the folding-unfolding strategy for three-fold problems might be too complex for some participants.

For relevant folds, the number of relevant folds (probe effect), the number of distractors implicitly depicting fewer than all relevant folds (distractor effect), and the interaction were significant (Figs. 11 and 12). This model outperformed the null, $\chi^2(3) = 10.14$, $p < .05$. Accuracy was highest for one relevant fold and lower for two and three relevant folds. Similarly, accuracy was highest when no folding-unfolding distractor was present and lower with one or two folding-unfolding distractors. For the interaction, the pattern matches that described for all folds.

Comparing the number of folds and number of relevant folds models, the number of relevant folds model did better, $\chi^2(0) = 2.11$, $p < .001$, revealing that participants likely ignore folds not relevant to the correct answer.

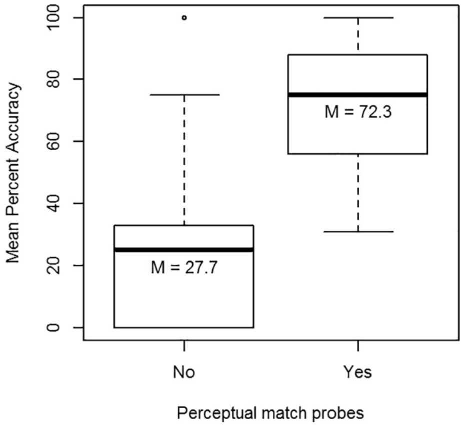

Fig. 13. Perceptual match: Accuracy predicted by perceptual match between the probe and correct item.

3.3.2. *Perceptual match*

This strategy was evaluated by examining whether the correct item had a punch in the same location as the probe (probe effect), the number of distractor items that also had a perceptual match (distractor effect), and the interaction. Only the probe effect was marginally significant (Fig. 13). Although this model did outperform the null model, $\chi^2(3) = 15.68$, $p < .001$ (Table C.2), we did not find much support for this strategy.

3.3.3. *Fold-to-punch strategies*

The **simple fold-to-punch** strategy was evaluated by noting whether the correct item contained two punches per probe fold (probe effect) and the number of distractor items that also had two punches per fold (distractor effect). The probe effect was not significant, but the distractor effect and interaction were (Fig. 14). This model outperformed the null, $\chi^2(3) = 16.54$, $p < .001$ (Table C.2), revealing participants may have used the simple fold-to-punch algorithm. The main effect of distractor number revealed a U-shaped function. Accuracy was highest with no or four simple fold-to-punch distractors, and accuracy was lowest when there were two such distractors (left panel of Fig. 14). When there are no simple fold-to-punch distractors, this algorithm quickly results in the correct answer. As the number of distractors that conform to the algorithm increase, so does the need to move away from using it. Performance is best when all distractors conform to the simple fold-to-punch algorithm as participants likely switch to another strategy.

The interaction effect revealed that when the correct answer did not match the simple fold-to-punch algorithm, having more simple fold-to-punch distractors decreased accuracy (Fig. 14, center panel). Participants seem drawn to simple fold-to-punch distractors. When the correct item matched the simple fold-to-punch algorithm, accuracy increased with the number of distractors (Fig. 14, right panel), also suggesting

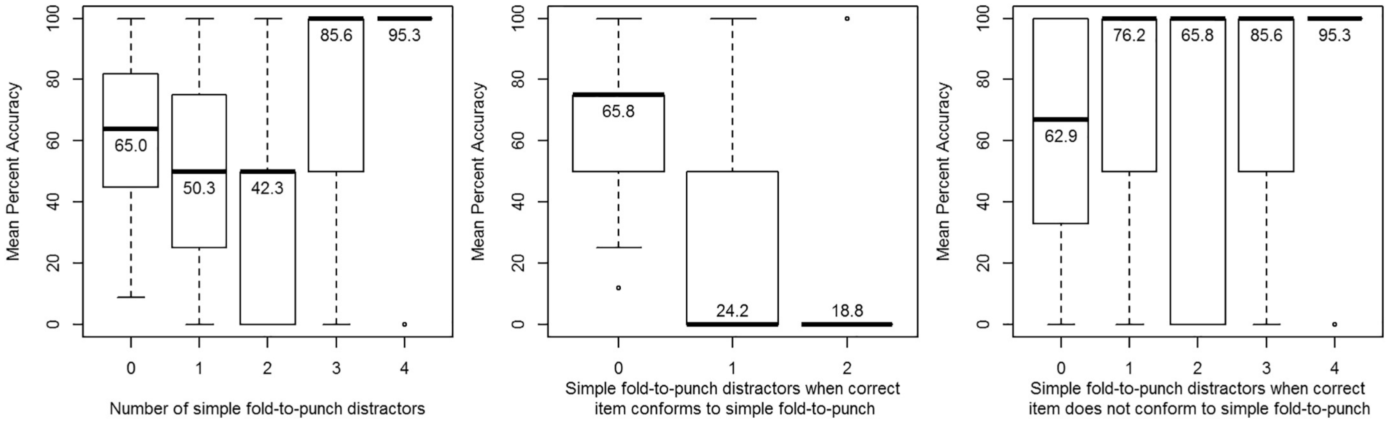

Fig. 14. Simple fold-to-punch: Accuracy predicted by the number of simple fold-to-punch distractors (left) and the interaction between when the correct and distractor items that follow the simple fold-to-punch algorithm (center and right).

participants switch to other strategies.

The **relevant fold-to-punch** strategy was evaluated using a model parallel to the simple fold-to-punch algorithm. This model did not outperform the null, $\chi^2(3) = 2.43$, $p = .49$.

3.3.4. Comparing the strategy models

Since the relevant folding-unfolding, perceptual match, and simple fold-to-punch models all outperformed the null model, we compared them. The perceptual match, $\chi^2(0) = 5.54$, $p < .001$, and simple fold-to-punch, $\chi^2(0) = 6.39$, $p < .001$, both outperformed the relevant folding-unfolding model. The simple fold-to-punch outperformed the perceptual match, $\chi^2(0) = 0.84$, $p < .001$. Although there is some evidence that participants used the relevant folding-unfolding and perceptual match strategies, overall accuracy better reflects the simple fold-to-punch algorithm.

3.4. Strategy categorization

A limitation of investigating strategy use using LMMs is that LMMs treat participants as if they all used the same strategy. Although that assumption is helpful in determining whether PFT accuracy can be explained by one strategy over another, it does not provide insights into individual differences in strategy use. Using the characteristics of the distractor items (Tables A.2 and A.3), we developed a categorization method to provide insights into strategy use (Appendix C).

The strategy categorization revealed that 73.1% of participants tended to favor one strategy: 22.1% folding-unfolding, 22.8% perceptual match, 10.1% simple, and 18.1% relevant fold-to-punch (Table C.1). One participant (0.7%) correctly answered all 20 items, so it was not possible to characterize the strategy this individual used. However, this participant likely used folding-unfolding because, if correctly applied, it always leads to the correct answer. The rest of the participants tended to favor multiple strategies: 20.8% favored two strategies and 5.4% favored three strategies.

4. Discussion

The PFT (Ekstrom et al., 1976) purportedly tests spatial visualization. Our results show that the PFT does not measure spatial visualization exclusively and it has the potential to reveal other strategies. We delved into how people solve PFT problems by investigating how accuracy relates to problem attributes and evidence for four strategies.

4.1. Problem attributes

Supporting our predictions, horizontal and vertical folds were the easiest. Diagonal folds were moderately difficult. Mixed fold problems (diagonal with vertical and/or horizontal) were the most difficult. These results support previous work showing that number of folds (Jaeger, 2015; Kyllonen et al., 1984) and fold combinations predict difficulty. However, few problems contained diagonal along with horizontal/vertical folds. In contrast with previous research (Jaeger, 2015; Kyllonen et al., 1984), occlusion type did not impact performance but few problems contained occlusions. A complete analysis of problem attributes would require an expanded set of PFT problems. By doing so, strategy use could also be better assessed.

4.2. Strategy use

Our models suggest that participants sometimes use the folding-unfolding strategy. The number of relevant folds impacted accuracy more than the total number of folds, suggesting that participants ignore irrelevant folds. Accuracy dropped with increasing distractors that accounted for fewer than the total number of folds. Thus, the PFT likely measures spatial visualization, to some extent.

A perceptual match between the probe and correct item also impacted accuracy, suggesting that the perceptual match strategy is a viable strategy. But perceptual match problems also tend to be easier than non-perceptual match problems (e.g., no occlusions, single fold types). We found limited evidence for this strategy.

We found good evidence for the simple fold-to-punch strategy, potentially in conjunction with other strategies. When the correct answer did not conform to the simple fold-to-punch strategy, participants tended to select distractors that did conform. This indicated that participants were likely using this strategy. In contrast, when the correct answer conformed to the simple fold-to-punch strategy, performance increased with more simple fold-to-punch distractors. In these cases, participants likely started solving a problem using this strategy, but then switched when multiple response items conformed to it. We found good evidence for this strategy, but more work is necessary to understand how this strategy interacts with other strategies.

When comparing models, the simple fold-to-punch outperformed the relevant folding-unfolding and perceptual match strategies. Although participants might have used multiple strategies, the simple fold-to-punch algorithm best predicted overall performance. Importantly, this suggests people do not solely engage in spatial visualization when completing the PFT.

As others have acknowledged (Glük, Machat, Jirasko, & Rollett, 2002), identifying strategies solely from item response models is problematic because it assumes that participants only use one strategy. To address this limitation, we developed a strategy categorization to reveal the strategies participants used. Participants tended to use the folding-unfolding and perceptual match strategies separately, followed closely by two strategies used together (most commonly folding-unfolding and perceptual match), then the relevant and simple fold-to-punch separately, and finally, very few participants used three strategies together.

Although our categorization revealed that most participants used one strategy, individuals can adjust strategies on a problem-by-problem basis (Just & Carpenter, 1985; Linn & Petersen, 1985). Jaeger (2015) found that participants reported other strategies along with imagining unfolding and that performance on difficult problems correlated with number of strategies reported. Because the PFT was not designed to evaluate strategy, the distractor items were not created to differentiate between the use of different strategies. As such, it was not possible for us to evaluate the strategies participants used on a problem-by-problem basis, using only the distractor items they selected. For strategy use to be determined for each PFT problem, there would need to be more problems and the distractor items would need to be systematically varied with respect to potential strategies. Future work could use other measures of strategy use to verify the efficacy of identifying strategies based on distractor items. Finally, neither of our strategy analyses could account for guessing. Future work should seek to either eliminate items that participants guessed on (using self-report or other metrics) or mathematically account for guessing.

4.3. Conclusion

Although the PFT has been broadly used to assess spatial visualization, our work suggests that people use spatial visualization but not exclusively. Our models suggest that participants use simple heuristics—like fold and punch counting—along with other strategies. Thus, rather than interpreting PFT performance as solely reflecting variation in spatial visualization, the PFT might also reveal variation in other approaches to paper folding problems. We would echo Just and Carpenter (1985, p. 170): “It would seem worthwhile to experimentally determine what stimulus characteristics govern the choice of strategy and then construct psychometric tests that systematically vary these characteristics.” The PFT falls short in its ability to evaluate strategies: it has too few problems and the distractor items were not systematically developed with this goal in mind. Therefore, we recommend developing a new PFT, one with more problems to allow for: more interactions between fold types, more occlusions, and more response items that conform to different strategies. This measure would capture variation in spatial visualization, when using that strategy, along with variation in the ability to use other strategies and shift between strategies.

Declarations of interest

None.

Acknowledgements

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305A140151 to Tufts University. The opinions expressed are those of the authors and do not represent view of the Institute or the U.S. Department of Education.

Appendix A. Attributes and strategies

Table A.1
Problem-level attributes: number of each type of fold, and type of fold occlusion.

Problem	Horizontal Folds	Vertical Folds	Corner Folds	Diagonal Folds	Fold Occlusion
1	0	0	1	0	None
2	1	1	0	0	None
3	1	1	0	0	Fold
4	0	0	0	2	None
5	1	0	1	0	None
6	0	1	2	0	Fold
7	1	0	0	1	None
8	0	0	2	1	Punch
9	0	1	0	1	Fold
10	0	0	1	1	Punch
11	0	0	0	1	None
12	1	1	0	0	None
13	2	0	0	0	None
14	0	0	0	2	None
15	0	0	1	1	None
16	2	1	0	0	None
17	1	0	1	0	Punch
18	1	0	1	1	Fold
19	0	1	2	0	None
20	0	1	0	1	None

Table A.2
Folding-unfolding strategy: number of folds in each probe and response items that aaccount for unfolding.

Problem	# Folds		Response Items that account for Unfolding					
	All	Relevant	A	B	C	D	E	# DIs
1	1	1	CI	0	0	0	0	0
2	2	2	0	1	0	CI	0	1

(continued on next page)

Table A.2 (*continued*)

Problem	# Folds		Response Items that account for Unfolding					
	All	Relevant	A	B	C	D	E	# DIs
3	2	1	0	CI	0	0	0	0
4	2	2	0	0	0	CI	0	0
5	2	2	1	CI	0	1	0	2
6	3	1	0	0	0	0	CI	0
7	2	2	CI	1	1	0	0	2
8	3	3	1	0	CI	0	0	1
9	2	2	0	0	0	1	CI	1
10	2	2	1	0	0	1	CI	2
11	1	1	0	0	CI	0	0	0
12	2	2	0	CI	1	1	0	2
13	2	2	CI	0	0	0	0	0
14	2	2	0	1	0	0	CI	1
15	2	2	1	CI	1	0	0	2
16	3	3	CI	2	0	0	0	1
17	2	2	0	0	0	1	CI	1
18	3	3	0	0	0	CI	0	0
19	3	2	0	0	1	CI	0	1
20	2	2	0	0	CI	1	0	1

CI: correct item; DI: distractor item.

Table A.3
Alternative strategies: number of correct and distractor items aligned with each strategy.

Problem	Answer	Perceptual Match		Simple Fold-to-Punch		Relevant Fold-to-Punch	
		CI	DI	CI	DI	CI	DI
1	A	Yes	BCDE	Yes	BCD	Yes	BCD
2	D	Yes	B	Yes	ACE	Yes	ACE
3	B	Yes	ACDE	No	–	Yes	A
4	D	Yes	ACE	Yes	BC	Yes	BC
5	B	Yes	ACD	Yes	–	Yes	–
6	E	Yes	CD	No	–	Yes	BCD
7	A	Yes	BCDE	Yes	–	Yes	–
8	C	Yes	AE	No	–	No	–
9	E	No	BCD	No	BC	Yes	AD
10	E	Yes	ABCD	No	–	No	–
11	C	Yes	AE	Yes	ABDE	Yes	ABDE
12	B	Yes	ACDE	Yes	–	Yes	–
13	A	Yes	DE	No	–	No	–
14	E	Yes	BC	Yes	A	Yes	A
15	B	Yes	ACE	No	E	No	E
16	A	Yes	BD	Yes	E	Yes	E
17	E	No	ABD	No	–	No	–
18	D	No	BC	No	–	Yes	ABCE
19	D	Yes	BC	No	–	Yes	AB
20	C	No	ABDE	No	A	Yes	BD

CI: correct item conforms to strategy; DI: distractor item conforms to strategy.

Appendix B. Linear mixed models

B.1. Attribute Problem LMMs

Using the “lme4” package in R version 3.1.2 (Bates et al., 2015), we developed a series of LMMs that analyzed how problem attributes (see Table A.1) affected accuracy on a problem-by-problem basis. The models included random effects for each participant and problem (null model). The following fixed effects were tested: number of folds (1–3), number of horizontal, vertical, corner, and diagonal folds (0–2 for each), occlusion (no occlusion, fold occlusion, punch occlusion; Table A.1), and presentation order. After the main effects were investigated, another model investigated how the number of each fold type interacted with each other. Each model was tested against the null model, and or tested against one another, using likelihood ratio chi-squares (χ^2). The estimates and standard errors for each fixed effect appear in Table B.1.

Table B.1
Estimates and standard errors for attribute linear mixed models.

Interaction effects model				
Intercept	1.40	0.25	5.52	***
Horizontal: Diagonal	−2.66	0.71	−3.75	***
Vertical: Diagonal	−3.29	0.71	−4.63	***

*p < .05.
**p < .01.
***p < .001.

B.2. Strategy LMMs

Using the “lme4” package in R version 3.1.2 (Bates et al., 2015), we developed four LMMs to investigate evidence for the four hypothesized strategies. Each model involved participants and problems as random effects with a probe effect, a distractor effect (Tables A.2 and A.3), and their interactions as fixed effects. Each strategy model had different probe and distractor effects, so they will be specified for each model. Each strategy model was run separately and tested against the null model, and or tested against one another, using likelihood ratio chi-squares (χ^2). Estimates and standard errors for each fixed effect are reported in Table B.2.

Table B.2
Estimates and standard errors for strategy linear mixed models.

	Estimate	*SE*	*z*	*p*
Folding-unfolding with total folds model				
Intercept	4.37	1.36	3.21	**
Probe	−1.63	0.64	−2.55	*
Distractor	−5.77	2.08	−2.78	**
Probe: Distractor	2.64	1.01	2.60	**
Folding-unfolding with relevant folds model				
Intercept	4.17	1.09	3.83	***
Probe	−1.86	0.61	−3.05	**
Distractor	−6.79	2.23	−3.05	**
Probe: Distractor	3.36	1.13	2.97	**
Perceptual match model				
Intercept	−1.51	2.26	−0.67	0.50
Probe	4.00	2.39	1.67	0.09
Distractor	0.06	0.73	0.08	0.93
Probe: Distractor	−0.48	0.78	−0.61	0.54
Simple fold-to-punch model				
Intercept	0.81	0.36	2.23	*
Probe	−0.16	0.61	−0.26	0.79
Distractor	−1.63	0.48	−3.41	***
Probe: Distractor	2.20	0.54	4.11	***

*p < .05.
**p < .01.
***p < .001.

Appendix C. Strategy categorization

A strategy categorization was developed to understand how participants differed in their use of the four strategies: folding-unfolding, perceptual match, simple and relevant fold-to-punch algorithms. Since correctly answered items problems reveal little about the strategies used, we focused on characteristics of the distractor items. To do this, we created scores for each of the four strategies, based on the properties of the distractor items (Tables A.2 and A.3) each participant selected. For instance, in solving a two-fold problem, an individual might make the mistake of selecting a distractor item that shows only one of the folds after unfolding. Selecting this distractor would suggest that the participant was using the folding-unfolding strategy (albeit unsuccessfully). In our categorization scheme, a participant would be given a score of 1 on this item problem for the folding-unfolding strategy. Since the PFT was not designed to assess strategy, some distractor items could suggest the use of multiple strategies, and other distractors do not suggest the use of any strategies. To address these issues, we dropped distractor items from the categorization that were not associated with any strategies (10/80 distractors), and we split the score evenly between multiple possible strategies. For example, item problem 1 distractors B, C, and D (Table A.3) could all be selected when using perceptual match, simple fold-to-punch, or relevant fold-to-punch. For this item problem, selecting any of these distractors would result in a score of 1/3 for perceptual match, 1/3 for simple fold-to-punch, and 1/3 for relevant fold-to-punch. Another issue is that the strategies have different numbers of distractors associated with them: folding-unfolding has 18, perceptual match has 56, simple has 18, and relevant has 29. The scores were equalized as if all strategies had 18 distractors. Participants were labeled as using one strategy over another if one strategy had 2.5 points (1 *SD*) more than any other strategy.

Table C.1
Number and percent of participants who were categorized as using the four strategies.

Folding-Unfolding		Perceptual Match		Simple Fold-to-punch		Relevant Fold-to-punch		All Correct	
33	22.1%	34	22.8%	15	10.1%	27	18.1%	1	0.7%

Folding-Unfolding with Perceptual Match		Folding-Unfolding with Simple Fold-to-Punch		Folding-Unfolding with Relevant Fold-to-Punch	
8	5.4%	1	0.7%	2	1.3%

Perceptual Match with Simple Fold-to-Punch		Perceptual Match with Relevant Fold-to-Punch		Simple with Relevant Fold-to-Punch	
5	3.4%	5	3.4%	10	6.7%

Folding-Unfolding, Perceptual Match, and Simple Fold-to-Punch		Folding-Unfolding, Perceptual Match, and Relevant Fold-to-Punch		Perceptual Match, Simple, and Relevant Fold-to-Punch	
3	2.0%	3	2.0%	2	1.3%

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Cooper, L. A. (1976). Individual differences in visual comparison processes. *Perception & Psychophysics, 19*(5), 433–444. https://doi.org/10.3758/BF03199404.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Gardony, A. L., Eddy, M. D., Brunyé, T. T., & Taylor, H. A. (2017). Cognitive strategies in the mental rotation task revealed by EEG spectral power. *Brain and Cognition, 118*, 1–18. https://doi.org/10.1016/j.bandc.2017.07.003.

Gardony, A. L., Taylor, H. A., & Brunyé, T. T. (2014). What does physical rotation reveal about mental rotation? *Psychological Science, 25*(2), 605–612. https://doi.org/10.1177/0956797613503174.

Glük, J., Machat, R., Jirasko, M., & Rollett, B. (2002). Training-related changes in solution strategy in a spatial test: An application of item response models. *Learning and Individual Differences, 13*(1), 1–22. https://doi.org/10.1016/S1041-6080(01)00042-5.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods, 48*(3), 829–842. https://doi.org/10.3758/s13428-015-0642-8.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence, 32*, 175–191. https://doi.org/10.1016/j.intell.2003.12.001.

Jaeger, A. (2015). *What does the punched holes task measure?* (Unpublished doctoral dissertation)Chicago IL: University of Illinois.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review, 92*(2), 137–172. https://doi.org/10.1037/0033-295X.92.2.137.

Khooshabeh, P., Hegarty, M., & Shipley, T. F. (2013). Individual differences in mental rotation. *Experimental Psychology, 60*(3), 164–171. https://doi.org/10.1027/1618-3169/a000184.

Kyllonen, P. C., Lohman, D. F., & Snow, R. E. (1984). Effects of aptitudes, strategy training, and task facets on spatial task performance. *Journal of Educational Psychology, 76*(1), 130–145. http://psycnet.apa.org/doi/10.1037/0022-0663.76.1.130.

Li, Y., & O'Boyle (2013). How sex and college major relate to mental rotation accuracy and preferred strategy: An electroencephalographic (EEG) investigation. *The Psychological Record, 63*(1), 27–42. https://doi.org/10.11133/j.tpr.2013.63.1.003.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*(6), 1479–1498.

Lohman, D. F. (1979). *Spatial ability: A review and re-analysis of the correlational literature (technical report no. 8).* Stanford, CA: Stanford University, School of Education, Aptitude Research Project (NTIS No. AD-A075972).

Shepard, R. N., & Feng, C. (1972). A chronometric study of mental paper folding. *Cognitive Psychology, 3*(2), 228–243.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701.

Wothke, W., & Zimowski, M. (1988). *Item analysis of the Paper Folding test (wks. 622). Technical report no. 1988-6.* Chicago, IL: Johnson O'Connor Research Foundation's Human Engineering Lab.